不确定性下的第三方惩罚：心理与脑网络机制*

李厅[1]，王迣[2]，罗跃嘉[3,4]，封春亮[5]

([1]四川师范大学脑与心理科学研究院，成都 610066)
([2]荆楚理工学院师范学院，荆门 448000)
([3]康复大学神经心理康复研究所，青岛 266113)
([4]北京师范大学认知神经科学与学习国家重点实验室，北京 100875)
([5]华南师范大学心理学院，广州 510631)

摘 要 第三方惩罚指的是当违规事件与自身利益无关时，第三方个体牺牲自我利益去惩罚违规者的行为。近年的研究提示，不确定性作为社会环境中一个重要且普遍存在的特征，可能是影响第三方惩罚执行的关键因素。然而，目前尚不清楚不确定性如何影响第三方惩罚及其认知与脑机制。本研究拟结合心理学与认知神经科学等跨学科技术，采用复杂脑网络前沿分析方法，（1）系统考察违规结果不确定性和违规意图不确定性对第三方惩罚的影响及其潜在的脑网络机制；（2）进一步探索不确定性驱使人们第三方惩罚行为改变的不同动机。本研究的成果不仅有助于从大尺度脑网络的整合角度加深对不确定性影响第三方惩罚机制的理解，还能为助推第三方规范维护行为等社会治理问题提供启发。

关键词 第三方惩罚，不确定性，公平，规范维护，脑网络

分类号 B842

1 问题提出

第三方惩罚(third-party punishment)指的是利益无关的第三方以损害自身利益为代价对违反了社会规范的个体所采取的制裁行为(Fehr & Fischbacher, 2004)。这一行为体现了“路见不平，拔刀相助”的正义精神，对于维护社会秩序至关重要(Fehr & Schurtenberger, 2018)，同时它也是心理学、行为经济学以及认知神经科学等领域的中心议题。

尽管实验室研究发现第三方惩罚在不同文化和年龄段的人群中普遍存在(Bernhard et al., 2006; Henrich et al., 2001)，但现场研究的结果并非如此(Balafoutas et al., 2014, 2016; Pedersen et al., 2020)。在现实生活中，旁观者的冷漠似乎也是一种常态。这种差异可能源于：旁观的第三方常常无法确定是否存在违规行为，或者该行为是否故意为之。与高度受控的实验室环境相比，生活中的决策往往基于不完整、模糊甚至模棱两可的信息，这种不确定性可能会妨

收稿日期：2024-12-16
* 国家自然科学基金青年项目(32300870)资助
通讯作者：封春亮，Email: chunliang.feng@m.scnu.edu.cn; 罗跃嘉，Email: luoyj@bnu.edu.cn

碍第三方惩罚的实施。然而，目前关于不确定性如何影响第三方惩罚的研究还相对较少(Toribio-Flórez et al., 2023)。

通常而言，违背公平的结果（即个体是否做出了不公平行为）和违背公平的意图（即个体是否故意做出不公平行为）是与第三方惩罚密切相关的两个因素(Buckholtz & Marois, 2012; Crockett et al., 2014; Krueger & Hoffman, 2016)。相应地，对不公平结果的不确定性和对违规者意图的不确定性都有可能影响第三方惩罚决策。近期的行为研究已经开始探讨结果不确定性对第三方惩罚的影响(Toribio-Flórez et al., 2023)，但鲜有涉及意图不确定性。考虑到人们在日常生活中频繁遭遇不确定性(FeldmanHall & Shenhav, 2019; Li et al., 2023)，系统地探究两种不确定性对第三方惩罚的影响，不仅能够丰富对规范维护心理机制的认识，还能够为助推第三方规范维护行为提供启发。

此外，第三方惩罚在结果不确定性和意图不确定性下的发生机制具体涉及哪些情绪和认知过程也尚不清楚，这限制了我们对规范决策加工的潜在机制的深入理解。在这方面，认知神经科学的发展为探索人类行为的基本认知结构提供了助力，且越来越多的共识是心理过程的映射更可能在于大尺度脑网络连接而非局部的脑区活动(Bassett & Sporns, 2017; Fornito & Bullmore, 2015)。尤其近期 *Science* 以特刊形式连续发表四篇综述，强调了脑网络连接是理解人类认知和行为的关键(Axer & Amount, 2022; Lee et al., 2022; Leergaard & Bjaalie, 2022; Thiebaut de Schotten & Forkel, 2022)。因此，解析结果不确定和意图不确定这两种情境下的第三方惩罚在脑网络层面的表征是重要且必要的。

总而言之，本研究拟结合心理学和认知神经科学等跨学科技术，采用脑网络前沿分析方法，以不确定性下第三方惩罚的心理与脑网络机制为主题展开系列研究。首先，分别考察结果不确定性和意图不确定性对第三方惩罚的影响，探究其潜在认知过程及脑网络基础。在此基础上，厘清人们的第三方惩罚行为在不确定情境下发生改变的不同动机，并探索不同动机类型者的脑网络模式差异。本研究将有助于完善对第三方惩罚内在动机的认识，促进从脑网络的整合角度理解人类利他惩罚的本质，以及为如何有效且高效地应对真实世界的违规行为提供新的见解。

2 国内外研究现状

2.1 公平背景下的第三方惩罚

公平是人类社会核心的价值追求，对于那些在资源分配过程中表现出自私的人，第三方会惩罚他们以维护公平。第三方惩罚研究的一个经典范式是基于独裁者游戏的第三方惩罚任务。在任务中，被试作为第三方，观看两个玩家进行独裁者游戏。其中一个玩家（独裁者）获得一笔金钱并决定如何在两人之间分配，而另一玩家（接受者）只能被动接受独裁者给出的分配方案(Kahneman et al., 1986)。当看到独裁者的不公平分配方案后，被试需要在公平维护和自我利益之间进行权衡，选择是否以付出自己的金钱为代价对独裁者进行惩罚(Fehr &

Fischbacher, 2004)。

当前，许多具有影响力的理论模型为我们揭示了人们在付出代价时仍愿意执行惩罚的心理动因(罗艺 等, 2013; 苏彦捷 等, 2019; 郑好 等, 2024)。最新的惩罚脑网络框架认为，在第三方惩罚的决策过程中，突显网络(salience network, SN)、默认网络(default mode network, DMN)及中央执行网络(central executive network, CEN)三大系统协同作用(Krueger & Hoffman, 2016)。首先，SN 涉及对规范违反行为的检测，引发愤怒、不公平厌恶等情绪反应；其次，DMN 受 SN 神经活动的调节，参与意图推理、评估等认知过程，并将其与违规结果进行整合，形成惩罚意愿；最后，CEN 将来自 DMN 的惩罚信号转换为实际的惩罚决策。前沿网络神经科学领域提出，解码人类心理过程的目标是探索脑区集合如何作为一个整体促进人类的认知与行为(Bassett & Sporns, 2017; Mišić & Sporns, 2016)。与此相呼应，这一框架突破了以往所关注的传统脑激活模式以及行为-脑单一映射关系，为我们全面理解第三方惩罚涉及的情绪和认知过程开辟了新途径(Li et al., 2022, 2024)。

尽管公平备受重视，但第三方惩罚决策也常受干扰。个体差异（如催产素水平、人格特质）与社会环境因素（如旁观者效应、框架效应）都会对此产生影响(吴燕, 周晓林, 2012; 周晓林 等, 2015; Guo et al., 2013, 2014; Zhu et al., 2019)。特别是，在现实生活中，第三方惩罚的发生频率并不高，也不一定随违规严重程度加剧。例如，在一项现场研究中，实验人员在德国科隆火车站这样的公共场合做出不同程度的违规举动，并统计了约 400 人次的近距离观察者反应。结果发现，仅约 20%的观察者实施了第三方惩罚，如对违规者表示谴责(Balafoutas et al., 2016)。同时，尽管观察者对较严重的违规行为表现出更强烈的负面情绪，并认为其应受到更严厉的惩罚，但实际上，严重违规行为与轻微违规行为的惩罚率并无显著差异(Balafoutas et al., 2016)。这或许是因为现实情境中信息往往不够充分，人们难以轻易判断违规的结果和意图。

2.2 不确定性对第三方惩罚的影响

社会环境总是充斥着各种不确定性，这对人类在社会互动中的思考与行动产生了深远影响。尽管在某些特定情况下，不确定性可能激发合作(Mill & Theelen, 2019)，但它更可能导致不符合群体利益的行为(FeldmanHall & Shenhav, 2019)。研究表明，人们对不确定性普遍感到厌恶，因为它会引发恐惧、焦虑和损失厌恶等负面情绪(Grupe & Nitschke, 2013; Tanovic et al., 2018)。在这些情绪的影响下，个体往往会更加关注自身利益和安全，亲社会行为（如互惠）也会减少(Guan et al., 2019; Vives & FeldmanHall, 2018)，甚至可能削弱道德推理能力(Kouchaki & Desai, 2015)。这意味着，在不确定性下人们在公平维护和自我利益的权衡中可能偏向后者，从而，第三方惩罚这类需要权衡两者的行为可能会减少。

目前，不确定性对第三方惩罚的影响已经取得了行为层面的证据。研究发现，当违规的结果是模糊的（如只知道分配方案的可能范围），第三方惩罚会减少(Toribio-Flórez et al.,

2023)。然而，与此相关的神经机制方面的研究尚且匮乏。近年来，关于不确定性如何影响亲社会行为的脑成像研究为相关脑机制提供了一些启示：处理他人帮助行为中的不确定性会激活与恐惧和焦虑等负性情绪处理相关的脑区，如眶额叶皮层和前侧脑岛(Xiong et al., 2020)；不确定性程度越高，与心理理论（如背内侧前额叶）和认知控制（如背侧前扣带回）相关的脑区活动也越强(Hu et al., 2017)。此外，信任决策（即在不确定的社会投资中所做的决策，这种投资既包含信任，也存在被背叛的风险）的脑网络研究发现，信任决策与情绪处理、心理理论以及认知控制相关网络之间的功能整合密切相关(Lu et al., 2019)。总的来说，这些发现表明，在社会领域中对不确定性的加工可能与情绪加工、心理理论和认知控制相关，但尚不清楚这些认知神经过程如何在不确定情境中影响第三方惩罚。

人们是否做出第三方惩罚的决策通常与当前情境中是否存在违背公平的结果和意图密切相关(Buckholtz & Marois, 2012; Crockett et al., 2014)，而对违规结果和意图的加工往往涉及不同的情绪和认知过程(Krueger & Hoffman, 2016)，从而可能导致截然不同的行为模式。例如，不公平结果会直接引发第三方惩罚(Fehr & Fischbacher, 2004)，并伴随着更高的自我报告的愤怒和前侧脑岛激活(Darley & Pittman, 2003; Harlé & Sanfey, 2012)。对不公平意图的加工则更多地涉及心理理论和认知控制过程，诱发了颞顶联合区、背外侧前额叶等脑区的活动(Feng et al., 2022; Güroğlu et al., 2010)。此时，第三方在决策过程中倾向于谨慎和宽恕，可能会牵制其做出惩罚的决定(Alter et al., 2007; Treadway et al., 2014)。此外，意图加工还可以通过调节人们对不公平结果的负性情绪反应来影响最终的惩罚决策(Yu et al., 2015)。因此，有理由推测，尽管两种不确定情境都可能导致第三方惩罚的改变，但在规范维护动机-利益权衡过程中可能存在不同的认知与神经机制，涉及不同的大脑系统之间的复杂交互。

2.3 不确定性下第三方惩罚改变的不同动机

如果结果和意图信息明确，第三方便可轻松解释当前情境，并考虑是否及如何对违规者采取行动(Malle, 2021)；当信息模糊时，人们在规范违反的解释上受到阻碍，从而影响第三方惩罚。以往研究表明，做出第三方惩罚这样的举动可以为人们带来声誉（如，值得信任）(Jordan et al., 2016; Raihani & Bshary, 2015)。在这里，声誉的获得是基于个体的公正决策，例如使“罪有应得”的违规者受到了制裁；然而，倘若制裁的是实际上并未违反或并非故意违反规范的人（即犯了第一类错误，“误伤好人”），则可能有损声誉或带来罪恶感。由此推测，人们在不确定性下可能会减少第三方惩罚以避免第一类错误。最近的行为研究证实了这一点：如果有机会减少不确定性，人们会采取主动措施（如获取更多信息）再决定是否惩罚；反之，人们不轻易惩罚(Toribio-Flórez et al., 2023)。这些结果突显了人们对公平规范的真正关心，而不确定性诱发的错误担忧可能是人们在规范维护动机-利益权衡中偏向自我利益、减少第三方惩罚的重要动机。

然而，最新研究表明，第三方惩罚可能并非那么情愿。具体而言，在经典第三方惩罚任

务的基础上，参与者得到了一个额外的机会，他们需要掷骰子并向实验人员汇报骰子点数的奇偶性，而掷骰子的结果决定了是否要执行自己刚才做出的惩罚决策。结果表明，最初做出惩罚决策的占 39.7%，随后，大多数人不诚实地报告了骰子的奇偶结果，导致最终的惩罚决策比例下降到了 8.6%(Kriss et al., 2016)。换句话说，有机会时，人们会谎报信息以避免实施第三方惩罚。此外，第三方惩罚通常被视为社会大多数成员对不公正行为的一种理所应当的回应(Fehr & Schurtenberger, 2018)。面对不公正行为选择无所作为，可能会被看作是纵容违规者，在一定程度上会损害个体的另一种道德声誉，进而影响其未来的社会参与(王博, 毕重增, 2021; Santos et al., 2011)。然而，不确定情境使其无法确认第三方是否在纵容违规。因此，不确定性减少第三方惩罚的原因可能还有利己的一面：人们可以利用不确定性来避免成本，同时避免承担纵容违规者等负面道德评价。类似地，在亲社会行为的研究中也发现，人们会策略性地将不确定性作为借口，为利己行为辩护或掩饰(Dana et al., 2007; Kappes et al., 2019)。由此可见，不确定性似乎可以成为一个绝佳的情境理由，让人们能够相对没有负担地选择自我利益，从而减少第三方惩罚。

综上所述，在预期不确定性减少第三方惩罚的背后，至少受到两种主要动机的驱使：避免第一类错误（即避免误伤好人）和避免成本。除此之外，在特殊情况下，不确定性也可能使第三方惩罚得以维持甚至增加。例如，最近一项基于社交媒体平台的研究发现，面对可能的不公正事件，常有人在违规信息未经证实之时，就义愤填膺地参与对事件或有损他人声誉的评论的传播(McLoughlin et al., 2024)。

2.4 发展动态分析与待解决问题

迄今为止，关于不确定性对第三方惩罚影响的研究已丰富了我们对社会互动复杂性的理解。然而，该领域的发展仍处于起步阶段，存在诸多亟待填补的空白。

首先，尽管在行为层面已初步证实不确定性对第三方惩罚的影响，相关的脑机制研究却相当匮乏。深入探究这一现象的脑机制至关重要，因为它能将大脑活动与决策过程中的不同认知过程（如情绪加工、意图推断）联系起来。特别是近年来，脑区间连接模式被视为人脑信息处理和认知行为的生理基础(Bassett & Sporns, 2017; Fornito & Bullmore, 2015; Thiebaut de Schotten & Forkel, 2022)。因此，从脑网络整合的角度考察不确定性对第三方惩罚的影响具有重要意义。

其次，现有研究主要集中在结果不确定性对第三方惩罚的影响上，而意图不确定性的影响尚未得到充分探讨。事实上，意图推断是执行第三方惩罚的关键因素之一。有些研究者甚至认为，公正的第三方更可能关注意图而非结果。例如，只有当违规行为是故意时，人们才会执行严厉的第三方惩罚(Treadway et al., 2014)。此外，结果和意图加工在惩罚决策中的情感投入存在差异：不公平结果会引发愤怒(Fehr & Fischbacher, 2004)，而意图推断时的态度则相对宽容(Alter et al., 2007; Treadway et al., 2014)。因此，有理由推测，在结果不确定性和意图

不确定性下，人们在自我利益与公平维护之间的权衡可能存在差异。通过比较这两种不确定性影响第三方惩罚的脑网络机制，有助于我们全面理解不确定性情境下的公平维护-利益权衡过程，从而深化我们对利他惩罚复杂性的认识。

第三，目前对于不确定性影响第三方惩罚的深层次动机鲜有探讨。如前所述，如果“避免第一类错误”和“避免成本”是人们在不确定情境下减少第三方惩罚的两种动机，那么也意味着不同动机的个体在对公平规范的维护上可能存在差异。基于此，我们首先通过提供消弥不确定性的机会（即观察个体在不确定性下是否寻求信息）来厘清不确定性如何影响人们第三方惩罚行为的改变，并进一步探究不同动机的个体是否可以通过即时（如加工不确定性时的脑网络组织模式）和/或稳定（如全脑灰质体积）的大脑网络特征进行预测。解答这些问题不仅有助于更准确地理解不确定性对第三方惩罚的影响，还能为个体规范维护动机的差异及其成因提供理论支持和潜在的脑干预策略。需要承认的是，不确定性对第三方惩罚的影响可能涉及多种动机和心理机制。本研究基于个体是否寻求信息的行为，采用简化的二元动机框架进行初步探讨，同时认识到这两种动机并非相互排斥，在特定情境下某种动机可能占据主导地位，从而影响个体的行为表现。

针对以上问题，本研究拟结合心理学、认知神经科学等跨学科方法以及复杂脑网络等前沿分析技术，系统考察结果不确定性和意图不确定性对第三方惩罚的影响及其脑网络机制，并阐明不确定性如何影响第三方惩罚行为的改变及其背后的不同动机；由此增进对人类规范维护行为的理解，鼓励更多个体参与社会规范的维护，并为社会公共秩序的宣传、教育及政策制定提供坚实的理论基础。

3 研究构想

3.1 研究 1 结果不确定性对于第三方惩罚的影响及其脑网络机制

实验内容：拟招募 30 名被试，性别比例均衡，年龄范围为 18~30 岁。被试需要在核磁内完成第三方惩罚任务的变式。该任务包含两种情境，共涉及三种具体的条件：结果不确定、结果不公平、结果公平。行为层面主要关注第三方惩罚在三种条件下的差异。脑机制层面，采用基于 SPM12 的 BASCO 工具箱实现任务诱发下的功能连通性的估计，并基于图论的方法，通过一系列定义明确、具有神经生物学意义且易于计算的度量标准来总结网络的系统级组织，从而刻画大脑是如何作为一个整体以支持结果不确定性下的规范维护-利益权衡，并进一步比较不同条件间的脑网络差异。

实验设计：采用第三方惩罚任务的变式，被试作为第三方，当看到独裁者的分配方案后，有机会对不公平事件采取行动。每个回合中，被试随机观看一对玩家（独裁者 *vs.* 接受者）完成独裁者游戏，独裁者决定如何分配一笔金钱，接受者只能接受。所有的分配方案都是由系统随机提出，但被试被引导认为玩家之间的互动都是真实的。在“结果确定”情境下（图 1a），每个回合开始时，独裁者决定了保留和给予的金钱数量，从而产生了一个明确的分配结

果，公平或不公平。依据平等原则(Deutsch, 1975)，本研究将 7:5 和 6:6 的分配提议定义为公平方案，并将 12:0、11:1、10:2、9:3、8:4 的提议定义为不公平方案。这种分类方法清晰、直观，是评估公平的常用手段(Fehr & Fischbacher, 2004; Fehr & Schmidt, 1999)。在看到独裁者的方案后，被试可以选择花费自己多少代币(monetary units, MUs)去减少独裁者的收益。每花费 1 个 MU，独裁者的收益减少 2 个 MUs(Fehr & Fischbacher, 2004; Wang et al., 2017)。最后，被试需要报告自己在当前回合的主观情绪。在“结果不确定”情境下（图 1b），第三方在屏幕上可以看到两个提议，其中一个是由独裁者此前做出的，另一个则由系统匹配以保证始终呈现一个公平提议和一个不公平提议，每个提议各有 50%的发生概率。这将导致一个模糊的结果，即第三方无法了解不公平分配是否确切地发生了。同样，在看到独裁者的方案后，被试可以选择花费多少 MUs 去减少独裁者收益，并报告自己在当前回合的主观情绪。

我们预期，相比结果不公平，人们在结果不确定下的第三方惩罚力度明显下降。对结果不确定情境的加工可能与情绪处理（突显网络）、心理理论（默认网络）、认知控制（中央执行网络）相关网络之间的功能整合有关。相比结果确定，结果不确定下突显网络的信息传递效率会更高效，可能反映需要消耗大量的心理资源去执行规范识别和情绪处理等过程。

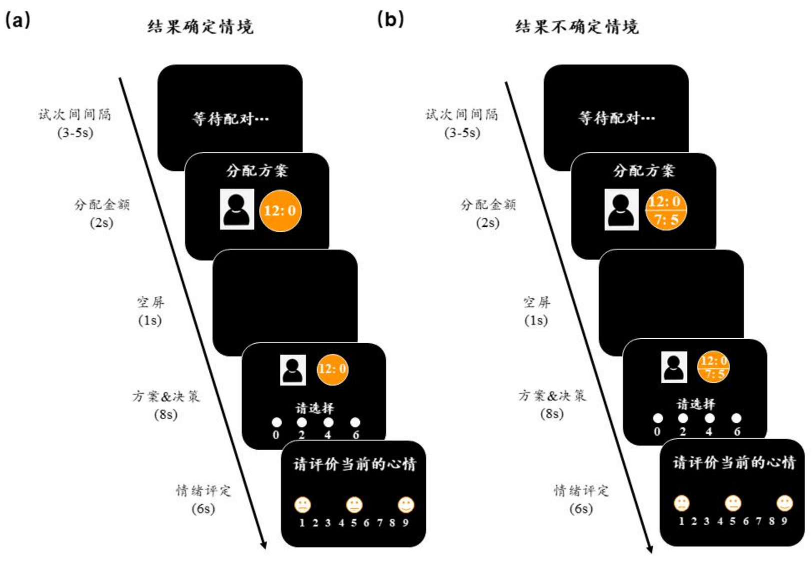


图 1　研究 1 流程图

3.2 研究 2　意图不确定性对于第三方惩罚的影响及其脑网络机制

实验内容：拟招募 30 名被试，性别比例均衡，年龄范围为 18~30 岁。被试需要在核磁内完成第三方惩罚任务的变式, 该任务包含两种不同的情境，共涉及三种具体的条件：意图不确定、意图不公平、意图公平。行为与脑机制层面的分析思路同研究 1。

实验设计：采用第三方惩罚任务的变式，被试作为第三方，当看到独裁者的分配方案后，有机会对不公平事件采取行动。每个回合中，被试随机地和多个玩家配对，观看他们完成独裁者游戏。所有的分配方案都是由系统随机提出，但告知被试玩家之间的互动都是真实的。在“意图确定”情境下（图 2a），每个回合由一个独裁者和一个接受者组成。独裁者需要在一

个公平和一个不公平分配方案中做出选择提供给接受者。被试作为第三方，在每一个回合开始时都将得到 6 MUs，看到独裁者的选择后，可以决定花费多少 MUs 去减少独裁者的收益。最后，被试需要报告自己在当前回合的主观情绪。在“意图不确定”情境下（图 2b），每个回合由两个独裁者和一个接受者组成。两个独裁者会看到一模一样的两个分配方案：一个公平（如自己保留 6，对方获得 6），一个不公平（如自己保留 9，对方获得 3），并在规定时间内独立做出选择。接下来，两个独裁者会被随机地选中一人（用红框表示）向接受者展示分配结果，未被选中者则与此轮分配的最后收益无关。只有当两个独裁者都选择公平方案时，该方案才有效；如果只有一人选择公平方案，或无人选择公平方案时，最后都会按照不公平方案来分配。也就是说，如果最后呈现的是一个不公平分配方案，那么该方案可能是被选出的独裁者的真实意图，也有可能是被另一个独裁者干扰的结果。同样，看到分配方案后，被试可以选择花费多少 MUs 去减少独裁者收益，并报告自己在当前回合的主观情绪。

我们预期，相比意图不公平，人们在意图不确定下对不公平行为的第三方惩罚力度明显降低。对意图不确定的加工可能与情绪处理（突显网络）、心理理论（默认网络）、认知控制（中央执行网络）相关网络之间的功能整合有关；相比意图不公平，意图不确定下默认网络的信息传递效率会更高效，可能反映需要消耗大量的心理资源去进行意图推断，并调节因不公平结果诱发的负性情绪体验。

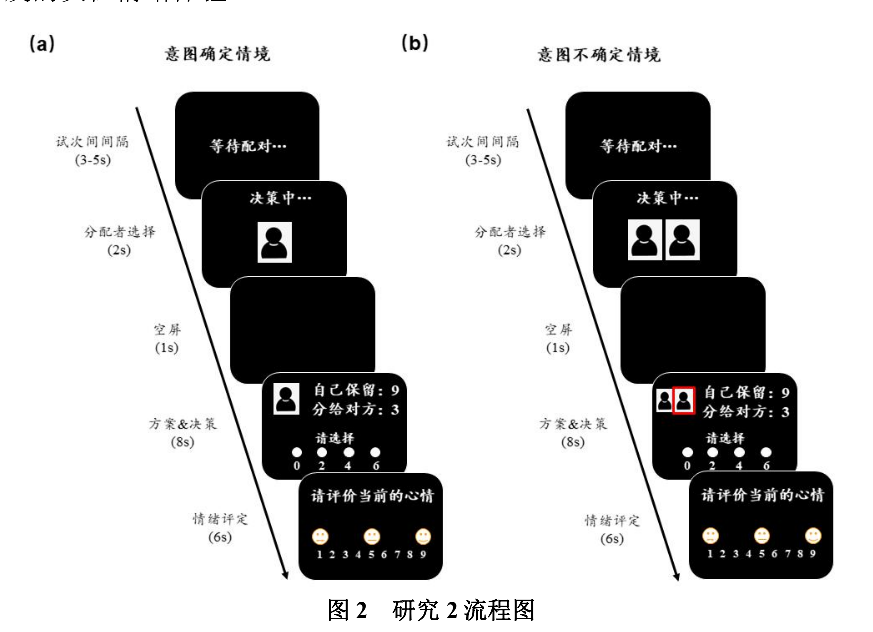


图 2 研究 2 流程图

3.3 研究 3 不确定性影响第三方惩罚的内在动机探析

实验内容：研究 3 分为两个阶段。首先，拟招募 200 名身心健康的成年志愿者完成行为任务。目的是在较大的样本中根据参与者的信息寻求行为，识别出由不同动机类型主导的个体。具体而言，如果参与者在面对结果和意图不确定性时，有超过 80%的试次选择了寻求信息，他们将被归类为“避免第一类错误”组；反之，如果超过 80%的试次未选择寻求信息，他们将被归类为“避免成本”组。行为任务结束后，从每个组中随机选取 30 名志愿者参与后续的

fMRI 扫描阶段。研究 3 在行为层面主要关注主动进行信息寻求的个体在不确定性被消除后第三方惩罚的改变。脑机制层面则关注即时的（加工不确定性时的脑网络组织模式）和/或稳定（灰质体积）大脑特征是否可以区分出避免一类错误和避免成本这两种动机类型者。功能脑网络构建同研究 1。结构脑网络的构建采用基于 SPM12 的 CAT12 工具箱对全脑灰质体积进行计算。

实验设计：行为任务在研究 1，2 的基础上，设置了解决不确定性的机会（图 3a）。以“结果不确定”为例，每一个回合中，看到独裁者的分配方案后，被试首先可以选择花费多少 MUs 去减少独裁者的收益，并报告自己在当前回合的主观情绪。接下来，被试被赋予机会应对不确定性或直接离开：如果选择应对，则结果不确定性被消除（也即确切金额被呈现）。随后，被试有机会调整自己的决策（被试知晓应对不确定性无需花费任何成本，且惩罚的实际效果以本次为准），并再次报告自己在当前回合的主观情绪；如果选择离开，则直接进入下一个回合。为防止由于道德声誉问题而产生的混淆效应，被试还被告知他们的个人信息和在每次试验中惩罚决定都不会被公开，由此，是否应对不确定性的选择可以区分出被试的不同动机。在“意图不确定”情境下完成同样的实验流程。两个任务的呈现顺序采用随机化的方式在被试间平衡。

间隔一周后，通过行为任务筛选出的两组被试（避免第一类错误和避免成本）再次来到实验室，完成第二个 fMRI 扫描阶段（图 3b）。在扫描仪中，参与者首先进行了结构像扫描，然后完成了与研究 1、2 相同的第三方惩罚任务。我们预期，不同动机类型者的大脑模式，特别是在奖赏加工、情绪处理和社会认知相关的脑区，显示出可区分的特征。

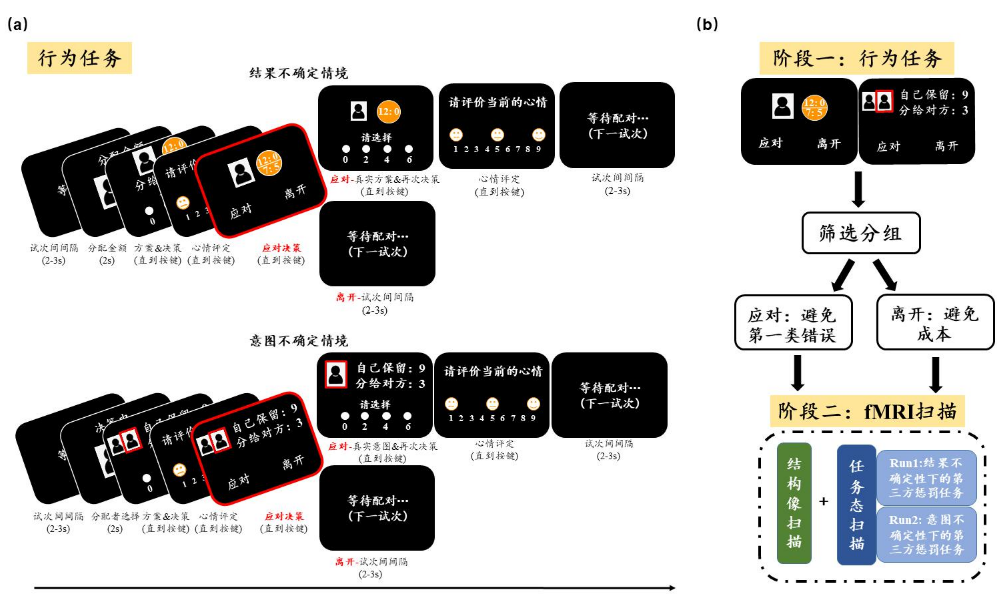


图 3　研究 3 流程图

4 理论建构与创新

日常生活中，人们在进行社会决策时，经常面临着信息不完全、结果不可预知、他人行

为难以预测等各种形式的不确定性。近年来的研究表明，不确定性往往削弱个体执行第三方惩罚的意愿(Toribio-Flórez et al., 2023)。作为一种维护社会规范的代表性行为，第三方惩罚的减少可能导致规范执行的稳定性下降，进而削弱规范的约束力，对社会整体利益产生不利影响。本研究综合运用心理学实验设计、fMRI 技术以及复杂脑网络分析等前沿方法，系统探讨不确定性如何影响第三方惩罚以及其背后的心理与脑网络机制。

在社会决策领域，以往的研究多集中在考察不确定性对分享、捐赠等亲社会行为的影响及其潜在心理机制(FeldmanHall & Shenhav, 2019; Li et al., 2023; Zhu et al., 2024)。第三方惩罚是一种典型的亲社会行为，因此也潜在地受到不确定性的影响。例如，在社会互动中，对行为结果的不确定（即该行为是否违反了社会规范）和对行为意图的不确定（即违规行为是否是故意的）都可以影响到人们的干预决策（即是否惩罚可能的违规者）。然而，直到最近才逐渐有研究关注结果不确定性与第三方惩罚的联系(Toribio-Flórez et al., 2023)。在本研究中，研究 1 和 2 同时探索了与社会决策紧密相关的这两种不确定性——结果不确定性和意图不确定性——对第三方惩罚的影响，并分别识别了在两种不确定性情境下，哪些情绪和认知过程的变化是影响第三方惩罚行为的关键因素。第三方惩罚包含了复杂的道德判断过程。根据道德判断领域的最新理论(Malle, 2021)，第三方惩罚与评价、规范判断、道德错误判断和责备判断等过程紧密相关。因此，通过探明哪些脑网络或脑区活动参与了与结果/意图不确定性下的公平决策相关的心理过程，并将这些活动与道德判断的不同过程相联系，本研究能够丰富我们对道德判断理论的认识，深化对个体在面对不同性质的不确定性时如何平衡维护如社会规范这样的群体利益与维护个人利益的认识，并为理解不确定性在更广泛的社会互动中的影响提供新的视角。我们预期，人们在结果和意图不确定情境下的第三方惩罚力度均下降，且自我报告的负性情绪增加。其中，突显网络（与情绪加工有关）的信息传递效率在结果不确定情境下更高效；默认网络（与意图推断/错误判断有关）的信息传递效率在意图不确定情境下更高效。

在研究 1、2 的基础上，研究 3 旨在进一步地厘清不确定性影响第三方惩罚行为变化的不同动机。更具体地说，不确定性对第三方惩罚行为的削弱可能与个体的动机差异有关。这种削弱在外部行为表现上难以区分，但其背后的神经机制可能存在显著差异。我们预期，腹侧纹状体、前侧脑岛、颞顶联合区等与奖赏加工、情绪处理和社会认知相关的脑区的拓扑属性（如节点度）能够区分避免第一类错误和避免成本这两种动机类型者。

在方法层面，本研究将应用复杂脑网络分析等前沿手段，解析不确定性影响下的第三方惩罚在情绪加工、意图推断等认知过程中的脑网络层面表征。复杂脑网络分析方法改变了传

统的将大脑视为数量巨大的离散解剖单元的研究方式，而是将大脑视为彼此相互连接的神经元构成的复杂统一体，继而关注并强调整个认知系统如何有组织的运行(Bassett & Sporns, 2017; Mišić & Sporns, 2016)。这是一场观念上的变革，为探索大脑内部工作机制提供了全新的视角，极大地扩展了当前社会规范决策领域相关研究的方法和思路。

总之，本研究旨在深入理解现实生活中不确定情况下的道德判断和决策，特别是那些与公平规范有关的情况，为公民道德教育体系建设、社会治安等方面提供新的见解。

参考文献

罗艺, 封春亮, 古若雷, 吴婷婷, 罗跃嘉. (2013). 社会决策中的公平准则及其神经机制. *心理科学进展*, *21*(2), 300−308.

苏彦捷, 谢东杰, 王笑楠. (2019). 认知控制在第三方惩罚中的作用. *心理科学进展*, *27*(8), 1331−1343.

王博, 毕重增. (2021). 道德声誉在第三方惩罚违规者行为认知中的作用. *中国社会心理学评论*, *21*, 153−165.

吴燕, 周晓林. (2012). 公平加工的情境依赖性: 来自 ERP 的证据. *心理学报*, *44*(6), 797−806.

郑好, 陈荣荣, 买晓琴. (2024). 第三方惩罚行为的认知神经机制. *心理科学进展*, *32*(2), 398−412.

周晓林, 胡捷, 彭璐. (2015). 社会情境影响公平感知及相关行为的神经机制. *心理与行为研究*, *13*(5), 591−598.

Alter, A.L., Kernochan, J., & Darley, J.M. (2007). Transgression wrongfulness outweighs its harmfulness as a determinant of sentence severity. *Law and Human Behavior*, *31*(4), 319−335.

Axer, M., & Amunts, K. (2022). Scale matters: The nested human connectome. *Science*, *378*(6619), 500−504.

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2014). Direct and indirect punishment among strangers in the field. *Proceedings of the National Academy of Sciences*, *111*(45), 15924−15927.

Balafoutas, L., Nikiforakis, N., & Rockenbach, B. (2016). Altruistic punishment does not increase with the severity of norm violations in the field. *Nature Communications*, *7*(1), 13327.

Bassett, D., & Sporns, O. (2017). Network neuroscience. *Nature Neuroscience*, *20*, 353−364.

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature*, *442*(7105), 912−915.

Buckholtz, JW., & Marois, R. (2012). The roots of modern justice: Cognitive and neural foundations of social norms and their enforcement. *Nature Neuroscience*, *15*(5), 655−661.

Crockett, MJ., Özdemir, Y., & Fehr, E. (2014). The value of vengeance and the demand for deterrence. *Journal of Experimental Psychology: General*, *143*(6), 2279−2286.

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory, 33*, 67−80.

Darley, J.M., & Pittman, T.S. (2003). The psychology of compensatory and retributive justice. *Personality and Social Psychology Review*, *7*(4), 324−336.

Deutsch, M. (1975). Equity, equality, and need: What determines which value will be used as the basis of distributive justice? *Journal of Social Issues*, *31*(3), 137−149.

FeldmanHall, O., & Shenhav, A. (2019). Resolving uncertainty in a social world. *Nature Human Behavior*, *3*(5), 426−435.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63−87.

Fehr, E., & Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behavior*, *2*(7), 458−468.

Feng, C., Yang, Q., Azem, L., Atanasova, K.M., Gu, R., Luo, W., Hoffman, M., Lis, S., & Krueger, F. (2022). An fMRI investigation of the intention-outcome interactions in second- and third-party punishment. *Brain Imaging Behavioral*, *16*(2), 715−727.

Fornito, A., & Bullmore, E. (2015). Connectomics: A new paradigm for understanding brain disease. *European Neuropsychopharmacology*, *25*, 733−748.

Guo, X., Zheng, L., Zhu, L., Li, J., Wang, Q., Dienes, Z., & Yang, Z. (2013). Increased neural responses to unfairness in a loss context. *Neuroimage*, *77*, 246−253.

Guo, X., Zheng, L., Cheng, X., Chen, M., Zhu, L., Li, J., Chen, L., & Yang, Z. (2014). Neural responses to unfairness and fairness depend on self-contribution to the income. *Social Cognitive and Affective Neuroscience*, *9*(10), 1498−1505.

Grupe, D.W., & Nitschke, J.B. (2013). Uncertainty and anticipation in anxiety: An integrated neurobiological and psychological perspective. *Nature Reviews Neuroscience*, *14*(7), 488−501.

Guan, F., Chen, J., Chen, O., Liu, L., & Zha, Y. (2019). Awe and prosocial tendency. *Current Psychology*, *38*(4), 1033−1041.

Güroğlu, B., van den Bos, W., Rombouts, S. A., & Crone, E. A. (2010). Unfair? It depends: Neural correlates of fairness in social context. *Social Cognitive and Affective Neuroscience*, *5*(4), 414−423.

Harlé, K.M., & Sanfey, A.G. (2012). Social economic decision-making across the lifespan: An fMRI investigation. *Neuropsychologia*, *50*(7), 1416−1424.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review*, *91*(2), 73−78.

Hu, J., Li, Y., Yin, Y., Blue, P.R., Yu, H., & Zhou, X. (2017). How do self-interest and other-need interact in the brain to determine altruistic behavior? *Neuroimage*, *157*, 598−611.

Jordan, J., Hoffman, M., Bloom, P., & Rand, D.G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature*, *530*(7591), 473−476.

Kahneman, D., Knetsch, J.L., & Thaler, R.H. (1986). Fairness and the assumptions of economics. *Journal of Business*, S285−S300.

Kappes, A., Nussberger, A.M., Siegel, J.Z., Rutledge, R.B., & Crockett, M.J. (2019). Social uncertainty is heterogeneous and sometimes valuable. *Nature Human Behaviour, 3*(8), 764.

Kouchaki, M., & Desai, S. D. (2015). Anxious, threatened, and also unethical: How anxiety makes individuals feel threatened and commit unethical acts. *Journal of Applied Psychology*, *100*(2), 360−375.

Kriss, P.H., Weber, R.A., & Xiao, E. (2016). Turning a blind eye, but not the other cheek: On the robustness of costly punishment. *Journal of Economic Behavior & Organization*, *128*, 159−177.

Krueger, F., & Hoffman M. (2016). The emerging neuroscience of third-party punishment. *Trends in Neurosciences*, *39*(8), 499−501.

Li, T., Feng, C. & Wang, J. (2024). Reconfiguration of the costly punishment network architectures in punishment decision-making. *Psychophysiology*, *61*(2), e14458

Li, T., Li, S., Li, B., Zhang, Z., Luo, Y., & Feng, C. (2023). Navigating uncertainty in human social decision - making: Consequences and coping strategies. *Social and Personality Psychology Compass, 17*(6), e12756.

Li, T., Yang, Y., Krueger, F., Feng, C., & Wang, J. (2022). Static and dynamic topological organizations of the costly punishment network predict individual differences in punishment propensity. *Cerebral Cortex*, *32*(18), 4012−4024.

Lee, J.H., Liu, Q., & Dadgar-Kiani, E. (2022). Solving brain circuit function and dysfunction with computational modeling and optogenetic fMRI. *Science*, *378*(6619), 493−499.

Leergaard, T.B., & Bjaalie, J.G. (2022). Atlas-based data integration for mapping the connections and architecture of the brain. *Science*, *378*(6619), 488−492.

Lu, X., Li, T., Xia, Z., Zhu, R., Wang, L., Luo, Y. J., Feng, C., & Krueger, F. (2019). Connectome-based model predicts individual differences in propensity to trust. *Human Brain Mapping*, *40*(6), 1942−1954.

Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology, 72*(1), 293−318.

McLoughlin, K. L., Brady, W. J., Goolsbee, A., Kaiser, B., Klonick, K., & Crockett, M. J. (2024). Misinformation exploits outrage to spread online. *Science*, *386*(6725), 991−996.

Mill, W., & Theelen, M.M. (2019). Social value orientation and group size uncertainty in public good dilemmas. *Journal of Behavioral and Experimental Economics*, *81*, 19−38.

Mišić, B., & Sporns, O. (2016). From regions to connections and networks: New bridges between brain and behavior. *Current Opinion in Neurobiology*, *40*, 1−7.

Pedersen, E.J., McAuliffe, W.H., Shah, Y., Tanaka, H., Ohtsubo, Y., & McCullough, M.E. (2020). When and why do third parties punish outside of the lab? A cross-cultural recall study. *Social Psychological and Personality Science*, *11*(6), 846−853.

Raihani, N.J., & Bshary, R. (2015). The reputation of punishers. *Trends in Ecology & Evolution*, *30*(2), 98−103.

Santos, M. d., Rankin, D. J., & Wedekind, C. (2011). The evolution of punishment through reputation. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1704), 371−377.

Tanovic, E., Gee, D.G., & Joormann, J. (2018). Intolerance of uncertainty: Neural and psychophysiological correlates of the perception of uncertainty as threatening. *Clinical Psychology Review*, *60*, 87−99.

Thiebaut de Schotten, M., & Forkel, S.J. (2022). The emergent properties of the connected brain. *Science*, *378*(6619), 505−510.

Toribio-Flórez, D., Saße, J., & Baumert, A. (2023). "Proof under reasonable doubt": Ambiguity of the norm violation as boundary condition of third-party punishment. *Personality and Social Psychology Bulletin*, *49*(3), 429−446.

Treadway, M.T., Buckholtz, J.W., Martin, J.W., Jan, K., Asplund, C.L., Ginther, M.R., Jones, O.D., & Marois, R. (2014). Corticolimbic gating of emotion-driven punishment. *Nature Neuroscience*, *17*(9), 1270.

Vives, M.L., & FeldmanHall, O. (2018). Tolerance to ambiguous uncertainty predicts prosocial behavior. *Nature Communications*, *9*(1), 2156.

Wang, L., Lu, X., Gu, R., Zhu, R., Xu, R., Broster, L. S., & Feng, C. (2017). Neural substrates of context-and person-dependent altruistic punishment. *Human Brain Mapping*, *38*(11), 5535−5550.

Xiong, W., Gao, X., He, Z., Yu, H., Liu, H., & Zhou, X. (2020). Affective evaluation of others' altruistic decisions under risk and ambiguity. *Neuroimage*, *218*, 116996.

Yu, H., Li, J., & Zhou, X. (2015). Neural substrates of intention-consequence integration and its impact on reactive punishment in interpersonal transgression. *The Journal of Neuroscience*, *35*(12), 4917–4925.

Zhu, R., Tang, H., Xue, J., Li, Y., Liang, Z., Wu, S., Su, S., & Liu, C. (2024). When advisors do not know what is best for advisees: Uncertainty inhibits advice giving. *PsyCh Journal*, *13*, 663−678.

Third-party punishment under uncertainty: psychological and brain network mechanisms

LI Ting[1], WANG Li[2], LUO Yuejia[3,4], FENG Chunliang[5]

([1]Institute of Brain and Psychological Sciences, Sichuan Normal University, Chengdu, 610066)
([2]Normal College, Jingchu University of Technology, Jingmen, 448000)
([3]Institute for Neuropsychological Rehabilitation, University of Health and Rehabilitation Sciences, Qingdao, 266113)
([4]State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, 100875)
([5]School of Psychology, South China Normal University, Guangzhou, 510631)

Abstract: Third-party punishment (TPP) refers to phenomena that unaffected bystanders punish norm violators at the cost of their own payoffs. According to recent studies, uncertainty—as an inherent feature of the social environment—may constitute a key factor modulating TPP behaviors. However, it remains largely unclear how uncertainty affects TPP as well as underlying cognitive and brain mechanisms. To address these issues, the current study adopts an interdisciplinary approach, bridging theoretical concepts and techniques across psychology, cognitive neuroscience, and the large-scale brain network analysis to (i) investigate the neurocognitive signatures of the effects of outcome or intention uncertainty on TPP; and (ii) explore different motivations driving the modulations of uncertainty on TPP. This study not only contributes to advancing the understanding of the role of uncertainty in TPP from an integrated perspective of large-scale brain networks, but also provides new insights into social governance, such as the design of institutions to uphold the social norms.

Keywords: third-party punishment, uncertainty, fairness, norm enforcement, brain network